Ecole Polytechnique Fédérale de Lausanne

MATH 342 - Time Series

Wisconsin Employment 1961-1975

Authors:
Vincent Micheli
Franck Dessimoz
Paul Jeha

Supervisor:
Anthony Davison

Date: 6th June 2019

Introduction

We analyse a time series of employment in various industries in Wisconsin from January 1961 to October 1975. The dataset consists of 178 monthly observations expressed in units of 1000 employees. It can be found in the tsdl R package (5).

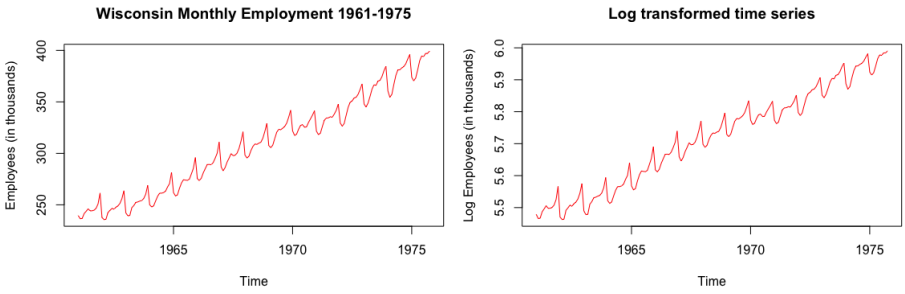
After performing an initial data analysis, we fit two classes of models to the data:

- SARIMA models
- Generalized additive models

Finally, each model is used to forecast the next two years of employment numbers.

The slide numbers in the following sections all refer to the lecture notes of (2).

1 Initial data analysis



(a) Time series (b) Log-transformed time series

Figure 1: Times series visualization

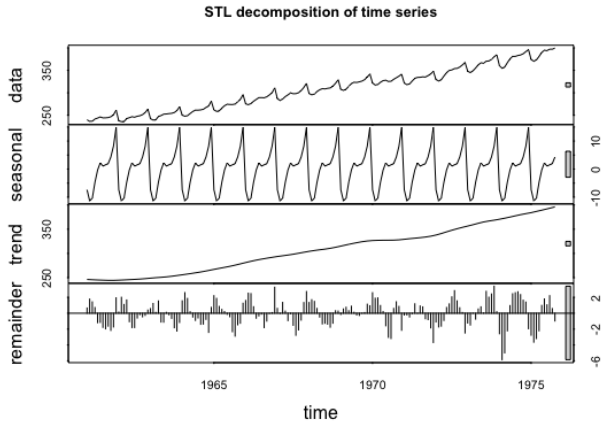


Figure 2: STL decomposition (Slide 106)

The time series (denoted as Y_t) is visualized in Figure 1 and 2. It shows an almost linear upward trend and periodicity of period 12 months. Moreover, the variance seems to be increasing over time. These observations indicate non-stationarity (Slide 24).

We log-transform the data in an effort to fix heteroskedasticity. This transformation is also suggested by Guerrero's automatic selection of the Box-Cox transformation parameter (3) (the method yields $\lambda = 0.02$). Therefore, in the following sections we model the log-transformed time series $W_t = log(Y_t)$ which does not show increasing variation with the level of the series.

2 Model fitting

2.1 SARIMA Models 1

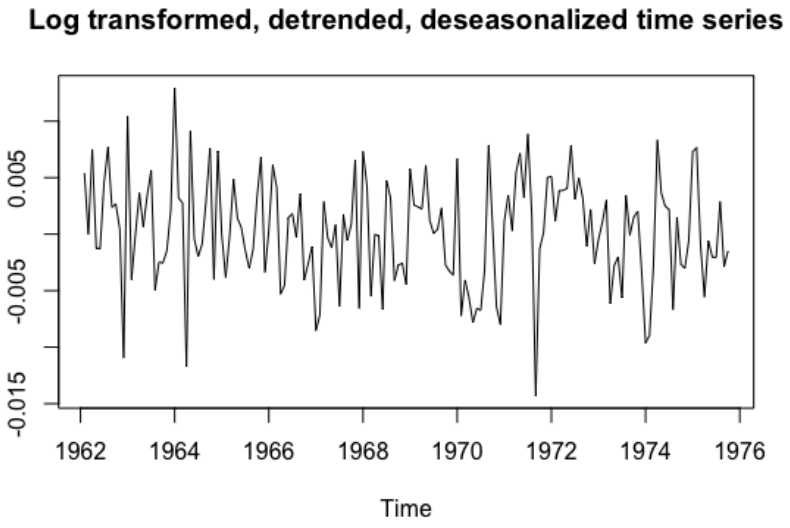
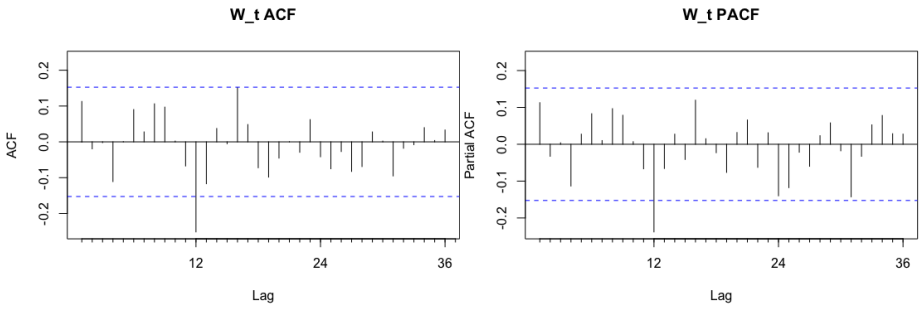


Figure 3: Stationarized time series

In order to model the time series as a SARIMA process (Slide 186), we follow the SARIMA modelling procedure explained in Slide 188.
We remove non-seasonal and seasonal trends by differencing (Slide 44) at lags 1 and 12. That is we consider $(1-B)(1-B^{12}) \times W_t$ shown in Figure 3 where B is the backshift operator. It appears we have obtained a stationary (Slide 65) time series (p-value of KPSS test greater than 0.1).



(a) ACF (b) PACF

Figure 4: ACF and PACF

	Coefficients	se
sma1	-0.327	0.081
σ^2	2.03×10^{-5}	
Log Likelihood	656.6	
AIC	-1309.2	

Table 1: SARIMA$(0,1,0) \times (0,1,1)_{12}$

	Coefficients	se
sma1	-0.314	0.083
sma2	-0.039	0.084
σ^2	2.027×10^{-5}	
Log Likelihood	656.7	
AIC	-1307.4	

Table 2: SARIMA$(0,1,0) \times (0,1,2)_{12}$

We fit and compare different SARIMA models based on their AIC (Slide 212) which trades model fit off against complexity. The correlogram (Slide 47) and partial correlogram (Slide 60) in Figure 4 suggest a SMA(1) component (Slide 187). Besides, the model given by the Box-Jenkins identification procedure yields the smallest AIC. Table 1 and 2 show the results obtained when fitting the SARIMA$(0,1,0) \times (0,1,1)_{12}$ and SARIMA$(0,1,0) \times (0,1,2)_{12}$ models.

Models	w	p-value
SARIMA$(0,1,0) \times (0,1,1)_{12}$ - SARIMA$(0,1,0) \times (0,1,2)_{12}$	0.21	0.65

Table 3: Likelihood ratio test

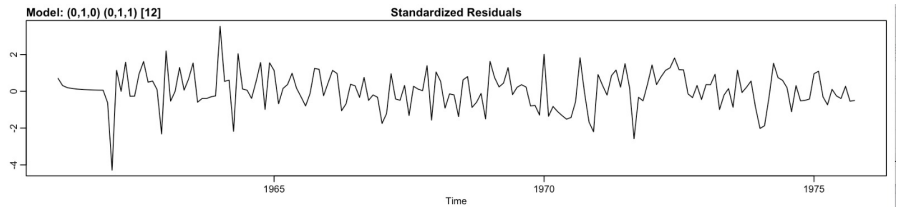
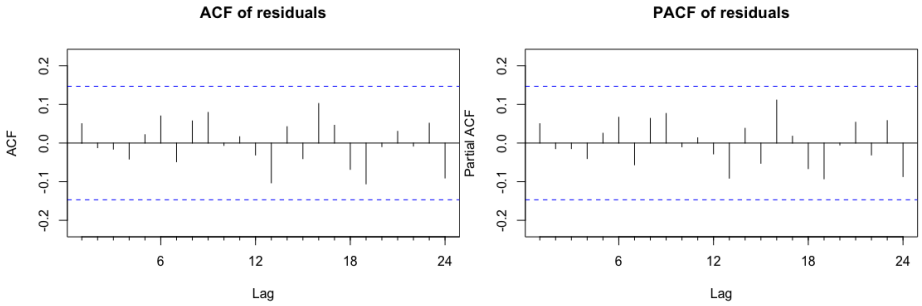


Figure 5: Standardised residuals

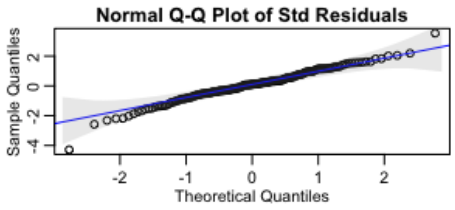


(a) ACF of residuals

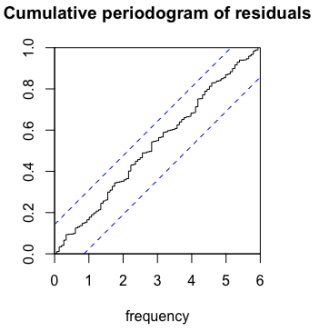
(b) PACF of residuals

Figure 6: Diagnostics 1

Table 3 shows a likelihood ratio test (Slide 133) comparing the two models. The p-value indicates we cannot reject the null hypothesis. That is, this parsimonious model is more adequate than a complex one. We still need to look at diagnostics to assess whether the model fits adequately or not.



(a) Normal Q-Q plot



(b) Cumulative periodogram

Figure 7: Diagnostic 2

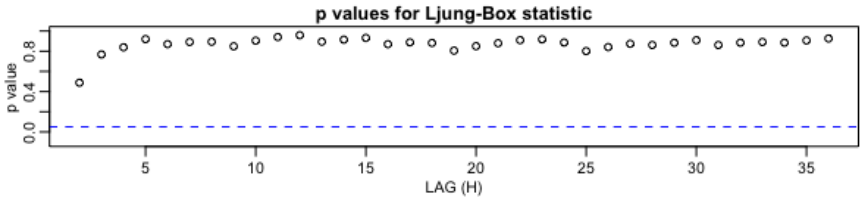
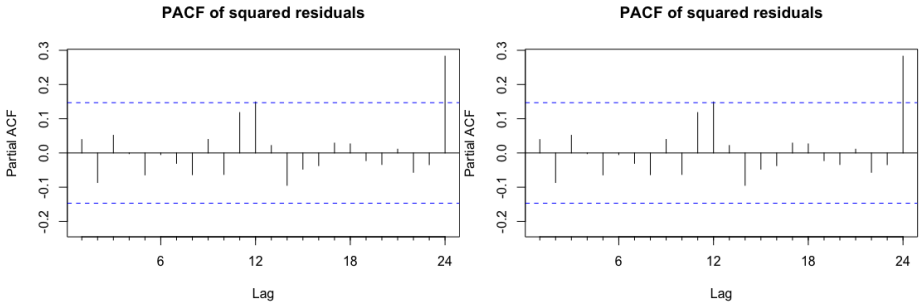


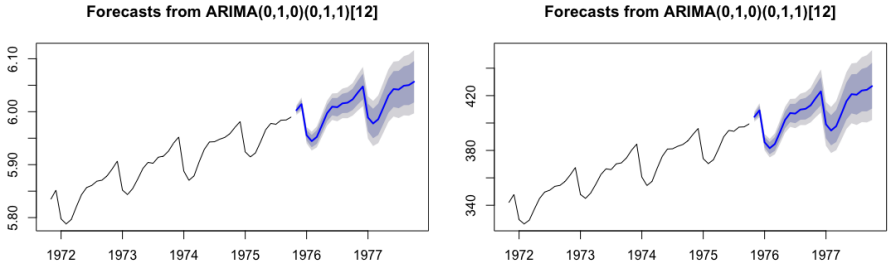
Figure 8: Diagnostics 3: p-values for Ljung box statistic

The ACF and PACF in Figure 6 do not show any autocorrelation structure in the residuals (Slide 143) even if we were to correct the confidence bands for small lags (Slide 173). The Normal Q-Q plot (Slide 67) of standardised residuals does not show any significant departure from normality except for one value outside the confidence bands. The cumulative periodogram (Slide 81) does not give any evidence against white noise. Both are shown in Figure 7. From the adjusted Ljung-Box test in Figure 8 (Slide 179), we fail to reject the null of linear independence of the residuals.



(a) ACF (b) PACF

Figure 9: ACF and PACF of squared residuals



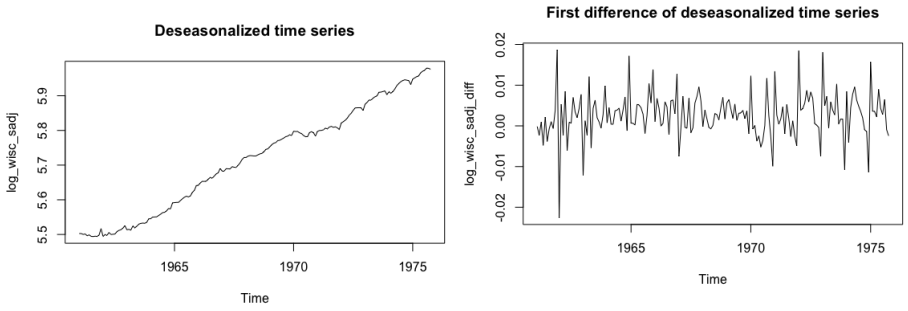
(a) Forecast on log-transformed data (b) Forecast on original scale

Figure 10: Forecasts (with 85% and 95% prediction intervals)

The ACF and PACF of squared residuals shown in Figure 9 do not suggest a SARIMA-GARCH model would be useful (Slide 222). From these diagnostic plots we cannot reject the idea that the standardised residuals are approximately Gaussian white noise.

Now that we have found a sensible model we forecast employment figures for the next two years (Slide 196). Moreover, the forecasts are back-transformed to the original scale in Figure 10. Even though we preserve the probability coverage of the prediction interval, it will no longer be symmetric around the point forecast (4).

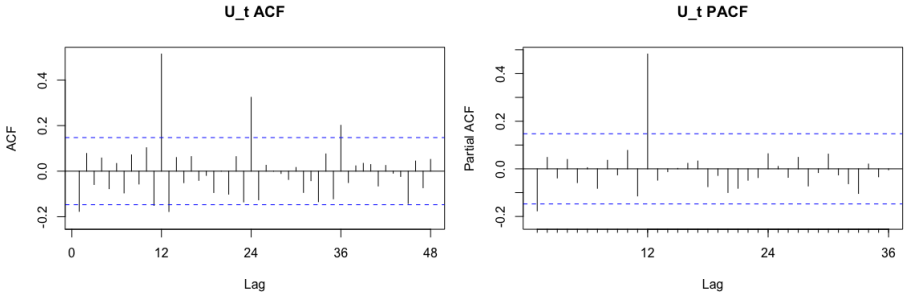
2.2 SARIMA Models 2



(a) Seasonally adjusted time series

(b) First difference of seasonally adjusted time series

Figure 11: Stationarization by removing a deterministic seasonal term



(a) ACF

(b) PACF

Figure 12: ACF and PACF

Now we perform an additive decomposition of the form $W_t = m_t + s_t + Z_t$ where m_t is the trend, s_t is a seasonal component and Z_t is a stochastic process. For a given month, s_t is the mean month value over the whole series. We remove the deterministic seasonal variation and fit a SARIMA model to the deseasonalized series. Figure 11 shows the first difference of this time series (denoted as U_t) which seems stationary (p-value of KPSS test greater than 0.1).

The ACF and PACF in Figure 12 mainly suggest a SAR(1) component (one could also look for AR/MA components). However, a $\text{SARIMA}(0,1,0)\times(1,0,1)_{12}$ model yields the smallest AIC as shown in Table 5.

	Coefficients	se
sar1	-0.719	0.054
σ^2	1.96×10^{-5}	
Log Likelihood	703.8	
AIC	-1403.6	

Table 4: SARIMA$(0,1,0) \times (1,0,0)_{12}$

	Coefficients	se
sar1	-0.828	-0.060
sma1	-0.222	0.104
σ^2	1.907×10^{-5}	
Log Likelihood	705.8	
AIC	-1405.6	

Table 5: SARIMA$(0,1,0) \times (1,0,1)_{12}$

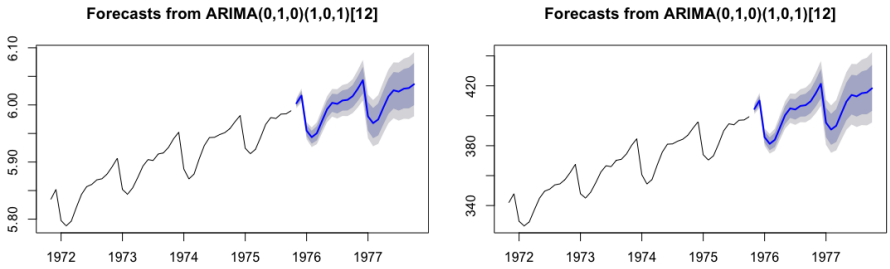
Models	w	p-value
SARIMA$(0,1,0) \times (1,0,0)$ - SARIMA$(0,1,0) \times (1,0,1)_{12}$	3.98	0.046

Table 6: Likelihood ratio test

The likelihood ratio test displayed in Table 6 indicates we should reject the hypothesis that the simpler model is adequate at level $\alpha = 0.05$.

The diagnostics plots (not shown) for the retained model have the same interpretation as in the previous section. Thus, we have no evidence against the idea that the standardised residuals are approximately Gaussian white noise.

Forecasts are then computed by including back the cyclic seasonal component. They are found in Figure 13.



(a) Forecast on log-transformed data

(b) Forecast on original scale

Figure 13: Forecasts (with 85% and 95% prediction intervals)

2.3 Generalized Additive Models

	edf	F	p-value
s(nMonth)	7.761	19.32	$< 2 \times 10^{-16}$
s(Time)	7.403	1957.47	$< 2 \times 10^{-16}$
R^2 (adj.)	0.988		

Table 7: Approximate significance of smooth terms

Once again we perform an additive decomposition but this time using a generalised additive model (1). That is, we model the data as

$$y = f_{seasonal}(month) + f_{trend}(time) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

where $f_{seasonal}$ and f_{trend} are smooth functions for the seasonal and trend features. We employ a cyclic cubic spline basis for the seasonal term in order to avoid discontinuities. No such restriction is put on the thin plate regression spline modelling the trend. This model choice is motivated by (6).

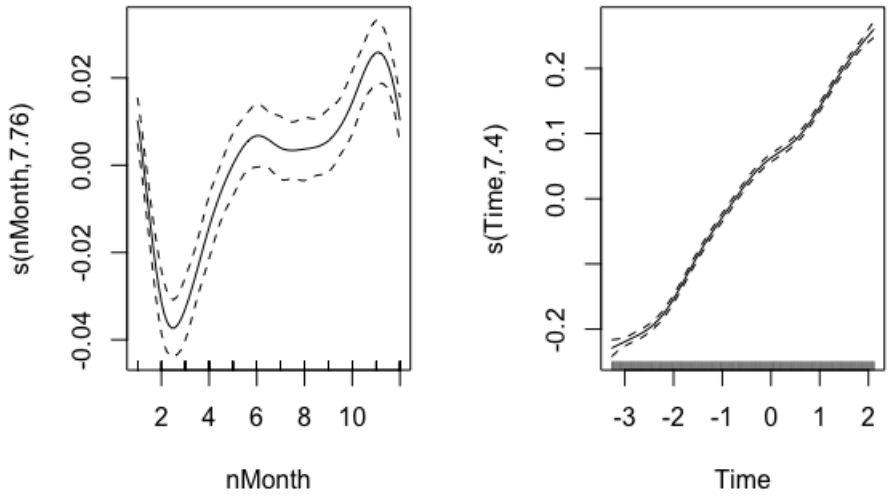


Figure 14: GAM smooth terms on log scale with 95% confidence bands

One could assume this model is spurious since it does not account for dependence in the data. This may result in smooth terms fitting noise to some degree. However, Table 7 shows that this model captures most of the variation in the data with highly significant smooth terms and an Adjusted R^2 close to 1. Nevertheless, we decide to model dependence within errors. Unfortunately the R package mgcv (7) does not allow users to fit models with a seasonal error structure. Therefore, we model this time series in a two-step process. First, we fit a generalised additive model assuming uncorrelated errors. Then we model the residuals as a SARIMA process.

The first step was already performed. Now we look at residuals. We remove non-seasonal and seasonal trends by differencing at lags 1 and 12 as shown in Figure 15.

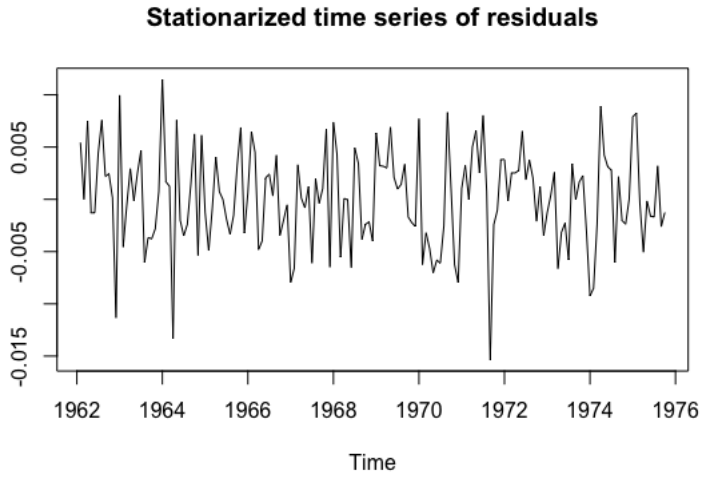
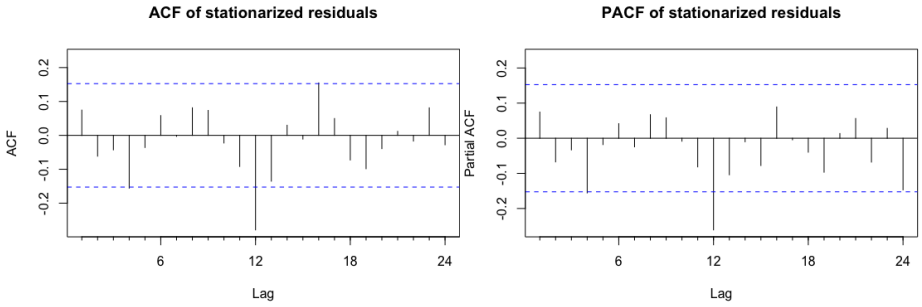


Figure 15: Stationarized residuals



(a) ACF (b) PACF

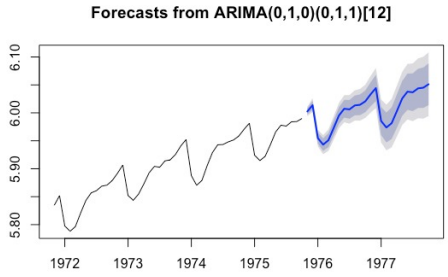
Figure 16: ACF and PACF of stationarized residuals

The ACF and PACF of stationarized residuals in Figure 16 suggest a SMA(1) component. This model is displayed in Table 7. Once again residual diagnostics (not shown) give no evidence against the hypothesis of Gaussian white noise.

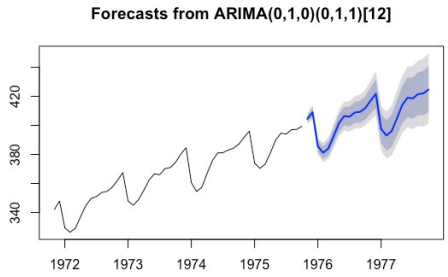
Forecasts on the log and original scales are then obtained by adding residuals forecasts to the two smooth terms previously computed. They are found in Figure 17. The prediction intervals displayed do not take into account the uncertainty associated to smooth terms.

	Coefficients	se
sma1	-0.3540	0.0764
σ^2	1.905×10^{-5}	
Log Likelihood	661.74	
AIC	-1319.47	

Table 8: SARIMA$(0,1,0) \times (0,1,1)_{12}$



(a) Forecast on log-transformed data



(b) Forecast on original scale

Figure 17: Forecasts (with 85% and 95% prediction intervals)

3 Conclusion

Many models were considered and fitted to this time series of employment. They differed in the way they handled trends in the data:

- The first approach was purely based on differencing and SARIMA modelling.
- The second one fitted a deterministic seasonal variation, removed it and then performed SARIMA modelling.
- The third one fitted a generalized additive model to the data and then performed SARIMA modelling on the residuals.

Each model provided a good fit while being parsimonious. In the end, forecasts were similar with a comparable level of incertitude. However, the generalized additive model may be a better fit for inference since it is the most interpretable model we considered. Indeed, it did not require differencing the data and consists in a summation of non-seasonal and seasonal terms.

References

[1] Davison, A. C. (2019) Modern regression methods lecture notes slide 155 onwards. Mathematics Section, EPFL.

[2] Davison, A. C. (2019) Time series lecture notes. Mathematics Section, EPFL.

[3] Guerrero, V.M. (1993) Time-series analysis supported by power transformations. Journal of Forecasting, 12, 37-48.

[4] Hyndman, R. J. and Athanasopoulos, G. (2018) Forecasting: principles and practice Chapter 3.5. OTexts.

[5] Hyndman, R. and Yang, Y. (2018) tsdl: Time Series Data Library.

[6] Simpson, G. (2014) Modelling seasonal data with GAMs. From the bottom of the heap.

[7] Wood, S. and Wood, M. S. (2015) Package 'mgcv'. R package version, 1-7.